DOCUMENT RESUME

ED 068 576 TM 002 088

AUTHOR	Carpenter, P.; Rapp, M.
TITLE	Testing in Innovative Programs.
INSTITUTION	Rand Corp., Santa Monica, Calif.
SPONS AGENCY	Department of Health, Education, and Welfare, Washington, D.C.
REPORT NO	P-4787
PUB DATE	Mar 72
NOTE	10p.
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Diagnostic Tests; Educational Innovation; *Instructional Innovation; *Measurement Techniques; Program Effectiveness; Remedial Instruction; Scoring; Teaching Procedures; *Test Construction; Test Validity

ABSTRACT

Some thoughts are presented on the appropriate amount of testing in innovative programs in education, the level of test difficulty, and the scoring and recording of test results. Most innovative programs require considerable testing; however, there is a danger of overtesting, which must be weighed against the desirability of obtaining information needed for a meaningful evaluation. Particular importance should be given to 1) selecting standardized norm-referenced tests of the appropriate difficulty in remedial programs; 2) selecting appropriate items for criterion-referenced tests, because their validity depended on accurate diagnostic testing; and 3) proper test administration procedures. (Author/LH)

TESTING IN INNOVATIVE PROGRAMS

P. Carpenter and M. Rapp

March 1972

P-4787

TESTING IN INNOVATIVE PROGRAMS

P. Carpenter and M. Rapp*

The Rand Corporation, Santa Monica, California

March 1972

*Any views expressed in this paper are those of the author. They should not be interpreted as reflecting the views of The Rand Corporation or the official opinion or policy of any of its governmental or private research sponsors. Papers are reproduced by The Rand Corporation as a courtesy to members of its staff.

The work on which this paper is based was performed pursuant to Contract No. HEW-OS-70-156 with the U.S. Department of Health, Education, and Welfare.

INTRODUCTION

Achievement tests are assuming a more important role than ever before in today's educational environment. They are being used to make decisions about the effectiveness of innovative programs and even to determine payments for performance contracts. Because of the importance being placed on test results, there is an urgent need to observe good testing practices. Test and measurement issues need to be considered simultaneously with the development of an evaluation design.

AMOUNT OF TESTING

Any innovative program is likely to involve a great deal of testing. Tests may be required for (1) diagnostic/prescriptive purposes, (2) determining mastery of assigned material, (3) determining the effectiveness of a program.

Many experimental programs stress individualization of instruction. By definition, this requires diagnostic testing and frequent checks of student progress. Most such programs, at a minimum, require pre- and post administrations of some standardized norm-referenced tests. In addition, in programs where criterion-referenced tests are used, periodic examinations to check mastery of objectives are required. Finally, the program evaluator may wish to administer some special tests. Moreover, unless the regular testing program of the district is suspended--and there are good reasons for not suspending it--these tests will also have to be administered.

The more tests administered, the less time is available for instruction and the greater the possibility of overtesting. Extensive testing may produce gains attributable to a practice effect, or it may diminish student interest in trying to do well on tests.

The appropriate response appears to be to keep the original test program design as sparse as possible in order to be able to add tests as their desirability becomes apparent and to minimize the chance of overtesting. Nonetheless, experimental programs require testing and past programs have been able to accommodate the demands. It appears wise to face up to the need to test in whatever amount is needed to obtain data to manage and evaluate the program.

LEVEL OF TEST DIFFICULTY

The choice of the proper level of test difficulty may have to be deferred until students have been chosen for the program, and even then some students may be given tests whose difficulty is inappropriate for them. This problem is extremely significant. If the test is too difficult for the student, his score may be due entirely to chance and the meaning of any computation of achievement gain will be questionable, regardless of how well other aspects of the test program or the instructional program itself are carried out. Such a situation is especially likely to arise when standardized, norm-referenced achievement tests are used because the tendency will be to give the test normed for the students' nominal grade level. But if one has a ninth-grade class, for example, selected on the basis that all students are at least two grade levels behind, a test normed for the ninth grade will produce only frustration and confused data. Picking a level of the test that reflects the actual achievement levels of the students, not their nominal grade placement, results in a more realistic picture of student achievement. However, this decision raises questions about how to interpret scores. It may even be necessary to administer tests of several levels of difficulty if the spread of student capability is very large. For example, ninth-grade students may be reading at all levels from pre-primer to the seventh grade. For children in lower grades with reading difficulties it may be extremely hard to find a written test with an appropriate level; some oral test may be required.

The problem raised by using the wrong level of test is underscored by the remarks of Ezra Mintz, executive vice president of Plan Education Centers, one of the contractors in the OEO performance contracting experiment. He told the Educational Marketer that "the results on which we were paid were garbage," because the test scores were inconsistent. In particular, Mintz said, in certain grades more than 30 percent of pupils tested actually regressed during the course of the year--raising the possibility that the wrong level of standardized test was administered. A spokesman for Alpha Learning Systems, another OEO contractor,

cited similar doubts, and said OEO had refused to furnish details on the tests used.*

There is a possible drawback to administering a test designed for a younger chronological age. A 12-year-old may only have the reading skills of the average 9-year-old but he will be interested in different topics. If the subject matter is inappropriate, it may interfere with the student's attentiveness to the point where he does not try to do his best.

Selection of appropriate tests of performance objectives (sometimes termed criterion-referenced tests) poses special problems. Their validity depends upon the accuracy of the diagnostic pretests. Performance objectives for each student are determined by diagnostic pretests. If these tests yield the correct set of objectives, a high score on the posttest indicates that the student mastered the objectives assigned to him and the test is, therefore, a measure of achievement. On the other hand, an invalid diagnostic test might result in performance objectives that are too low for the student, so that a high test score would not indicate any real achievement.

PREPARATION OF TESTERS

In past innovative programs tests have been given by trained psychometrists, counselors, teachers, or graduate students. In general, it would seem that the qualifications of the tester are less important than are the preparation and training given the tester and the tester's sense of responsibility for assuring the reliability of test results. In addition, whether the tester is a local person or from outside the district, advance preparation is needed so that arrangements can be made. Otherwise, unnecessary confusion and bad temper may result. Coordination is needed between the administrators of the program and the school principal or building administrator responsible for the rest of the school's academic program so that times and places for testing can be arranged to cause a minimum of disruption to both the regular program and the innovation.

---

*The Educational Marketer, White Plains, New York, Vol. 4, No. 10, p. 2.

Before administering the tests, the tester should be familiar with the facilities in which the tests will be given, the number of students to be tested, the times available for testing, and procedures for obtaining, safeguarding, and collecting test booklets and for scoring and recording test results. This is especially important if the test administrator is from outside the district, as he will be unlikely to understand the local ground rules.

Each tester should be provided with a manual and a sample copy of the test several days before the examination and urged to study the manual and to practice by taking the test himself. He should also be provided with a written set of instructions outlining his duties at all stages of the test. Tests of performance objectives may require that testers practice giving the directions and reading the questions (if given orally). If test items require students to work with actual objects, such as a dictionary, testers may need instructions for judging whether student responses are correct.

If the test is timed, testers should be supplied with timers (such as a stopwatch or an interval timer) that will provide the accuracy required; they should not be expected to provide their own. They should be instructed to give the students exactly the time designated to complete the test, whether they seem to have done all they can in less time or seem to be pressed for time. Otherwise, it is almost impossible to interpret the results, particularly in the case of standardized tests.

CONDITIONS OF TEST ADMINISTRATION

Writers of standardized tests assume a set of conditions and procedures for test administration. These assumed conditions are described in all good testing texts and are well known to most teachers and administrators. In fact, however, school districts seldom comply fully with the principles. Giving tests under standardized ideal conditions is costly and time consuming and, in many school buildings, meeting the standards comes close to being infeasible. The problem is even more severe for tests of performance objectives, many of which require the tester to observe the student as he undertakes some task.

If proper testing conditions are not provided, student scores may be contaminated by factors unrelated to their actual achievement. Tests of performance objectives that are to be attained by students in the program should also be administered under conditions that are as close to those specified in the texts as possible. Proper test conditions have been determined after years of expert experience with testing, and they apply equally to standardized tests and to tests of performance objectives, even though the latter are still in the developmental stage. On the other hand, achieving the proper textbook-required conditions may be infeasible or so difficult that to achieve the ideal conditions would be prohibitively expensive.

This dilemma is probably best met head-on by an explicit recognition that examinations are not likely to be given under antiseptic textbook conditions. Therefore, just how the tests are to be administered becomes an important consideration in planning the evaluation.

It is unfortunate that schools whose primary achievement measures are scores on tests rarely have the proper facilities for administering tests under the conditions specified by most test constructors. Most schoolrooms today provide adequately for the comfort of students although some are still overly hot in late spring or early fall. Probably the most serious concern is space between students sufficient to discourage copying. Although there may be deliberate attempts to cheat by some students, perhaps of more concern is the opportunity for inadvertent copying or simply the distractions when students are seated so close together that they can see someone else's answers. Students of above-average ability are likely to have enough confidence in their answers so that they will not change them because they see that someone else has a different answer. Students of low ability, on the other hand, are rarely confident of their answers and are much more likely to be influenced by their neighbor's answers. In a case like this, where we know that many responses are in the chance range, everything possible should be done to reduce opportunities for copying.

The requirements for space will depend, to a degree, on the type of answer the test calls for. Standardized tests are usually answered by making small marks on an answer sheet, and these are difficult to

see at a distance of more than a few feet. If the average classroom is built to accommodate about thirty students, then only about fifteen should be tested at a time by a standardized test. If performance objectives are being tested, student responses will be more varied, and may often be easy to see or hear even from a considerable distance. In this event, it is possible that carrels should be used to remove distractions or opportunities for copying.

Because of class schedules, it is difficult to cope with the problem of providing good testing conditions. It may, for example, be necessary to increase the time the evaluator must devote to testing in order to administer tests properly in existing facilities. It could be money well spent if it improves the reliability of the test scores.

DIRECTIONS

If students do not understand what they are supposed to do, test data are invalid. Perhaps the most difficult problem to remedy occurs when the vocabulary of the directions is too advanced for the students on the low end of the distribution. The directions of the tests may have to be adapted to the vocabulary level of the population being tested. On the other hand, this conflicts with the standardization of directions for norm-referenced testing. It is a problem that should be discussed during the planning of the evaluation. A decision can be made about whether complying with standardization procedures or obtaining more precise results is more important.

If students are not already familiar with the test directions and the format of the answer sheet, a practice session should precede the actual testing. This is routine procedure on such standardized tests as the Scholastic Aptitude Test and the Graduate Record Examination which provide students with samples of questions before they take the tests.

Information necessary for proper identification and recording of test results must frequently be entered on answer sheets or the front of test booklets. The tester must insure that this information is accurate and complete, either by entering it himself or by instructing the students on how to do so. In the latter event, he should spot-check

student entries before the examination begins so that needed data will not be lost.

We have been dwelling on "nuts-and-bolts" aspects of examinations even though most educators have received training in testing. In many actual testing situations textbook-specified or test-instruction specified procedures have not been followed. More important, in actual testing situations we have observed we questioned whether test instructions and procedures elicited the behavioral responses desired by the testers and test authors. Perhaps many students failed to understand the mechanics of test-taking.

MONITORING TEST ADMINISTRATION

If the program has an auditor he can forestall possible questions about testing conditions by observing during their administration. He can determine, for example, if time limits are properly adhered to, if directions are being given uniformly by all administrators, and if the physical conditions surrounding the test are the best possible under the particular constraints of any district. Furthermore, he can obtain information of value to the evaluator by observing whether students seem to be following directions or answering at random. This kind of observation can be useful in making a judgment about the appropriateness of the test for the population to which it was administered.

An evaluator can perform the same functions if the program does not have an auditor. If the evaluator is responsible for test administration, however, this puts him in the position of certifying his own work.

SCORING AND RECORDING TEST RESULTS

Most achievement tests are machine scored, so that there is little need for the evaluator to check the accuracy of scoring. If he has confidence in the test bureau doing the scoring he will have confidence in their having conducted the appropriate checks. In the lower grades, where students mark the test booklet directly, a random check of scoring accuracy should be made to determine whether all booklets should be

rechecked. Tests of performance objectives will probably also be scored manually. If so, at least a random check of the scores should be made by someone who did not perform the initial scoring. If such checks show discrepancies, it may be necessary for all tests to be scored twice by independent scorers.

When tests are scored by machine, the scoring and recording is performed simultaneously. This means that answer sheets must carry all of the necessary identifiers so that the scores may be recorded in the proper format. Much time that would otherwise be spent in wasted data-handling can be saved by making sure that answer sheets are properly filled out. Again, a random preliminary check, followed if necessary by complete checking may be the best procedure to use. Another option is to pass out answer sheets the day before the test and have students fill in the required information under a teacher's supervision. When the sheets are collected, they can be double-checked and missing information supplied before the answer sheets are distributed to the students at the testing session.

SUMMARY

The nature of most innovative programs requires considerable testing and adds importance to maintaining appropriate procedures. The danger of overtesting needs to be weighed against the desirability of obtaining the information needed for a meaningful evaluation.

A particularly important issue is how to select standardized norm-referenced tests of the appropriate difficulty in remedial programs. Selection of appropriate items for criterion-referenced tests is critical because validity depends on accurate diagnostic testing.

It is important that proper test administration procedures be enforced. The physical conditions of most schools and the pressures of daily school-life make this difficult, and some testing practices in past programs have raised questions about the meaningfulness of the results.